Date, Time & Location:	**Vocabulary and Common Data Elements** **Face-to-Face Meeting Notes** **July 19, 2004, 8 am-6 pm ET**
Attendees:	UPMC • Rebecca Crowley • John Gilbertson (TBPT Liaison) Jackson Lab • Jim Kadin • Martin Ringwald University of Hawaii • Lynne Wilkens • Leo Cheung Mayo • Chris Chute • Harold Solbrig Albert Einstein • Xin Zheng Fred Hutchinson • Bob Robbins • Dan Geraghty UC-Davis • Cecil Lynch Ohio State University • Scott Oster (Architecture Liaison) Patient Advocate • Ben Rude EMMES Corporation • Brian Campell • Claudine Valmonte

	NCICB • Peter Covitz • Frank Hartel • Denise Warzel NCI • Larry Wright • Margaret Haber • Chitra Mohla SAIC • Kathleen Gundry • Tommie Curtis BAH Team • Christine Richardson • Michael Keller • Mark Adams • Theo Wills • Doug Tidquist • Patty Disandro • Greg Eley
Agenda Item #1:	**I. Vocabulary Development Governance Model** Moderator: Frank Hartel • Frank Hartel introduced the topic of vocabulary and the need to develop on governance model for caBIG. The governance model for controlled vocabularies may not (and probably will not) be the same as the CDE governance model. • The following issues were also raised by Frank Hartel: o Selection Process as a group for terminology resources o Interactions with publishers o Standardization with existing resources o Deployment of those resources o Influence future development o Extension of these resources

- Center/server ratio
- Shared caBIG Vocabulary server
- Mix of terminology embedded in each center

- Peter Covitz expressed the need to document the interaction of caBIG participants with external organizations possibly by filling out a form to submit to the Strategic Planning Working Group.
- Might not really be that strategic because groups will still be able to interact with external organizations without documenting or reporting it. V-CDE WS is responsible for control of this.
- Bob Robbins stated that V-CDE governance models cannot be at ends with caBIG. He also expressed the following concerns:
  - Institution infrastructure vs. Community infrastructure
    - Ability to optimize in the community, but can not bind it
  - Scope Creep
  - Query computer results vs. Information Retrieval
  - Quality of data associated with grid
    - Ideally, the quality of data on the grid will be as if you put it there yourself.
  - Syntactic vs. Semantic Interoperability
    - Both will evolve.
    - Caveat that groups with rich semantics won't want to participate if forced because the system is not to that level of sophistication.
- Three potential governance models for Vocabulary:
  - Centralized Governance
    - caBIG can have an impact on the community by securing positions on editorial board, as reviewers and in general, by getting involved.
    - Martin Ringwald asked if a common model is needed in all cases or if vocabulary governance can be addressed as a case-by-case situation to determine the best course of action

for that particular case.

- The model itself is self-consistent.
- Chris Chute agreed and stated that CDEs can function as a widget between the information model and vocabularies.
- It is necessary to work closely with the Architecture WS to make sure of consistency.
- No one terminology will meet the needs of caBIG.

- Federation (Loose)
  - Two terminologies exist (Terminology A and B)
    - No common schema
    - Semantic website
    - NCI Thesaurus and MGED Ontology are 2 examples
    - Can pick appropriate terminology from each
    - Critiques communicated manually (i.e. by phone)
    - Current state of affairs
- Federation (Formal)
  - Untried
  - Does caBIG forsee this model?

- Harold Solbrig stated that there are intermediate points between these three options.
  - For example, leave LOINC in its own model and space and reference when appropriate. This would fall between loose and formal federation.
- How would terminologies be consumed that are not specifically federated?
  - An extension of the centralized model would be required.
- Concern was expressed about caBIG becoming a central authority of terminologies in a group of existing central authorities.

- Syntax interoperability is the first step in getting people to realize they have semantic interoperability.
- It appears that there is preliminary consensus that a federated model would work best with caBIG.
    - One vocabulary would never work.
    - There is a need to allow freedom of different definitions.
- There are currently three different terminology resources in caBIG.
    - NCI develops one major terminology resource.
    - Two of the Cancer Centers will be developing terminology resources.
    - There are also additional external standards the V-CDE WS will agree to include for use in caBIG.
- There must be sensitivity to the distinction between individual terminologies and the mapping between those terminologies.
- The vocabulary governance model has to be tied to the deployment model (i.e. tied to architecture).
- Local changes in terminology must be available immediately in the whole system.
- **Summary/Action Items:**
    - The V-CDE WS must decide on a practical governance model for vocabularies in the next couple months.
    - This model will not necessarily be the same as the CDE governance model.
    - Three main models are:
        - Centralized
        - Federation (Loose)
        - Federation (Formal).
    - The governance model may actually be an intermediate point between the three options.

II. Selecting Standards for caBIG

Moderator: Kathleen Gundry

- Kathleen Gundry provided a brief introduction to the External Data Standards Review that was prepared for

NCICB by SAIC. This document is a living document. Kathleen proposed the following questions for the V-CDE WS to address.

  - How will caBIG choose standards to apply universally?
  - Are there standards already in use that are so compelling we will use them?
  - Essentially, what is the process for selection of standards?
  - How will we IT-enable them?
  - How do you apply standards?
    - Not in favor of standards police

- The WS felt that this document covers the spectrum of content and interchange standards.
- An additional question arose concerning how these standards will meet/bind/and be used in the interoperability we envision for caBIG?
- **Information Processing Model**
  - Chris Chute submitted that caBIG is as much an information model as anything else.
  - This is assuming the information model is at the application layer.
  - The Information model needs to be coupled with data standards to constrain aspects of the model.
  - Do we need to adopt a formal modeling structure like RIM?
  - Clinical applications need to be HL7 compliant and undergo modeling appropriately.
  - Is that a necessary solution?
    - Structured Data Objects are not as abstract as RIM.
    - This achieves absolute levels of interoperability.
    - This may be necessary for CTMS, but not necessarily throughout caBIG.
  - What is the long-term strategic view?

- Vocabulary models pertain to information model.
- We need to address the need for formal information models/semantic interoperability models.
- CDEs by their nature, have a clearly defined role between the Information Processing Model and vocabularies.

- How do we deal with changes in terminology and meaning say over a ten-year period?
  - Federating over time or space?
  - Using CDEs as a coupler between vocabulary and information model gives flexibility in the evolution of both.
- What is the consensus on an information model driving vocabulary development?
  - Each Domain WS would have to provide input; would not be top down.
  - Need to have class/object model built around application or domain
- Data is being gathered in a variety of ways and different standards are going to be required to reflect that.
- What do standards apply to?
  - All surface interfaces (Data In/Data Out) rather than all internal components
  - Must declare supported standards at interface and map to it unambiguously with metadata associated
- HL7 was designed to provide interface.
  - Problems will arise if System A has a different information model from System B.

- **Summary/Action Items**
  - External Standards Document needs to be distributed.
  - Liaisons to the Domain WS need to go to respective WS and gather standards that are used.
  - Need to establish an Information Processing Model
  - Not only do caBIG participants need to use external

standards, but they also have to announce what standards they are using.

- Using CDEs as a link between Vocabularies and Information Processing Model allows for evolution in both.
  - VCDE must develop a set of guidelines for relationships between vocabs, CDEs and data models
- Versioning of standards must also occur.

III. CDE Curation and Administration

Moderators: Kathleen Gundry and Denise Warzel

- Kathleen Gundry and Denise Warzel provided a brief overview to CDE curation and administration.
  - The current operation in existence for clients is still missing the broader notion of information modeling.
  - caDSR currently interacts with the NCI Thesaurus; if it is not in the Thesaurus then it looks in the Meta-Thesaurus.
  - UML model-driven CDEs
    - Class diagrams transformed into metadata
- **caDSR Training**
  - Beta training done
  - Who should be trained?
  - What should the training material be?
  - How to use tools?
  - Where should the training occur?
    - Webcasts, On-site "where gathered"
  - When should training begin?
    - Looking at September
  - Denise Warzel proposed that training should be in reverse
    - Start with V-CDE WS
    - Train in Designated Domain WS
    - Then to Cancer Centers

- A modification was suggested to the above mentioned flow of training.
  - Instead of training the entire V-CDE WS, train the Domain WS liaisons and the respective V-CDE liaisons.
- APIs, Browser, Administration and Curation Tools
  - Which are most crucial?
  - Should be demand-driven
  - May not need to be just caDSR, also EVS or object model development
  - Rather teach processes than tools
  - How do you keep CDEs consistent and up to date/implement and maintain in applications over time?
  - Internal group developed strategy that any upgrade of system archives them at each upgrade
- Training must include why we are standardizing up front.
  - We need training session/material that needs to include an overview because people are getting the machinery without knowing why or what it is for.
- In order to determine how closely things are related, we need to understand why we are working with these things.
- Need concrete examples of how CDEs have helped, as well as limitations
- By binding CDE with semantics in the vocabulary system, you can leverage metadata, thesaurus and meta-thesaurus.
- Modeling data provenance
  - We are looking to the Architecture WS to determine what provenance modules should be covered.
- EDRN is a tissue database that is published in a deliberate de-identified manner that is an example of something using CDEs.

- **Summary/Actions Items**
  - The overriding use case for CDEs is to be able to use other people's data without knowing each other and being able to rely on that data.
  - Reference implementations should include analysis of interoperability which will drive metadata.
  - Training requirements need to be identified.
  - The appropriate Domain WS liaisons and V-CDE members also need to be identified to receive the training.

IV. Perspectives on CDE Curation

- Brian Campell from EMMES Corporation provided an presentation and overview of CDEs from the CTEP project.
- Questions that arose from the presentation included:
  - How many CTEP CDEs are actually active?
    - 2300 are released and approved for trial use; a lot have been retired and phased out.
  - What differences or commonalities do CDEs have for different types of cancer?
  - How do you deal with merging CDE's across studies?
- No formal connection between caDSR and EVS
  - CRF question/CDE question relationship
  - Want to leverage the richness of the ISO 11179 model to the information model
- Tommie Curtis next gave a brief presentation and overview of CDE use with DCP and CCR and in relation to C3D.
  - There is a need to agree on the appropriate role and scope of the information model, CDEs, and vocabularies.
  - The challenge is that different communities have different points of reference.

V. caBIG Compatibility Document

Moderators: Chris Chute and Harold Solbrig

- Harold Solbrig provided an overview of the caBIG compatibility document to the WS. The floor was then

opened to everyone for comments on the document.

- This document is driven by demand.
- Boundaries need to be defined more crisply.
- Need to map uses and look where use case breaks on information model and figure out how to un-break it
- Security and Confidentiality needs to be addressed.
- Can not have disjointed levels of compatibility (i.e. Silver in one area, bronze in another).
- Is there anything missing in the middle two rows?
    - If a standard is used, then that standard needs to be declared.
        - CDEs will make that declaration.
    - Standards change with versions.
- There is a bias in the document toward source applications; need to address conduit and sink applications.
- Who is blessed authority to review applications for caBIG compatibility?
- RDF is preferable to expose metadata for information model.
- Semantic architecture is within the scope of V-CDE.
    - The Information model will ultimately get instantiated into Architecture, but guidelines on how to use UML models is V-CDE's responsibility.
    - It is very desirable to include Information model in the scope of V-CDE WS because all WS will be dealing with this to share a high level information model across caBIG.
- Can we come up with a mechanism to build a model that can use existing data classes and be modified as needed?
- Think about idea of platinum level of compatibility with the addition of consent/permission management.
    - Clinical data objects carry consent information: portable/global permissions model

- Any server of data objects should have key in perpetuity

- **Summary/Action Items**
  - Requirements for caBIG security, confidentiality
  - Immutable identifiers and other aspects of inter-resource referential integrity
  - Compatibility for data analysis, transformation and presentation applications and user applications

VI. Managing Survey Metadata

- Dan Gillman from the Bureau of Labor Statistics gave a presentation on managing survey metadata including CDE development and management.

VII. Use Cases

- Various members of the V-CDE WS presented various Use Cases to the meeting participants.
  - University of Hawaii use case
    - Lynne Wilkens and Leo Cheung presented a use case discussing how dietary exposure is used in epidemiology of cancer.
    - Do we need more/less different information?
    - They need assistance in understanding the NCI thesaurus and meta-thesaurus to become caBIG-compatible.
  - Jackson Lab use case
    - Jim Kadin and Martin Ringwald presented use cases dealing with MGI mouse anatomy.
    - Questions that were presented by Jim Kadin:
      - In framing use cases to structure vocabulary, how do we get enough detail for vocabulary?
        - Use cases are for problem solving and work flow.
      - How do you deal with non-specific anatomy?
      - How do we structure these use cases appropriately for pathology vs. expression?

        - Define stakeholders for vocabulary.
    - What is a vocabulary versus a primary data object?

- Albert Einstein use case
    - Xin Zheng presented a use case on Albert Einstein's bioinformatics shared resource.

- UC-Davis use case
    - Cecil Lynch presented a use case dealing with adverse event reporting-CTCAEv3.
    - Need cross referencing ICD-9, MEdDRA, SNOMED

- UPMC use case
    - Rebecca Crowley presented a use case that involves developing standards for caTISSUE with manual and automated annotation of pathology information.
    - caTIES is a text information extraction system.
    - How do we tie these annotation CDEs into the information model?
        - Not yet mapped into the information model

- NCI/OC use case
    - Larry Wright presented a use case on automated coding of pathology reports.
    - It is very important to map to concepts and between concepts

- **Summary/Action Items**
    - Use cases will predominantly be from Domain WS to V-CDE WS for terminology issues.
    - Explicitly add who the stakeholders are in the use cases from the Domain WS.
    - Collect Domain WS use case requirements and analysis before joint Face-to-Face meeting with Architecture.
    - Domain WS leads need to be notified to get background and stakeholder use case for vocabulary and metadata.

Action Items:	**Name Responsible**	**Action Item**	**Date Due**	**Notes**
	All	Identify a practical, functional vocabulary governance model.		
		Deployment of governance model		
		Develop (make up) a set of guidelines for relationships between vocabs, CDEs and data models.		
		Documentation of the Use of a standard		
		Domain-related standards		
		Versioning of deployed standards		
		Training requirements and prioritization of groups that need training		
		Requirements for caBIG security, confidentiality		
		Immutable identifiers and other aspects of inter-resource referential integrity		
		Compatibility for data analysis, transformation and presentation applications and user applications		
		Address conduit and sink applications in compatibility document		

[redacted]		Addition of stakeholders to use case format		
[redacted]		Starting condition of vocabularies or metadata to be included in background		
[redacted]		Communicate with Domain WS leads to gather background and stakeholder use cases for vocabulary and metadata.		

Please list below and attach Meeting Materials and Agenda (if prepared separately):